# Data Specification for Machine Learning Project

Contents

[Data Specification for Machine Learning Project 1](#_Toc185794136)

[Data Requirements 2](#_Toc185794137)

[Data Dictionary 3](#_Toc185794138)

[Data Quality 4](#_Toc185794139)

[IDA/EDA Summary 5](#_Toc185794140)

[Preprocessing 6](#_Toc185794141)

[Data Security and Privacy 7](#_Toc185794142)

## Data Requirements

Dataset overview

|  |  |
| --- | --- |
| Data source |  |
| Data format | Structured |
| Data size | Xx records  Xx features |
| Data type | CSV |
| Data storage | SQL database |
| Update frequency | Monthly |

Data Fields

Target Variable

Xxx or xxx

Features – Econ & Social data

* Feature values represent proportion of people in different xx type (index) having done something (column)
* e.g. below means 55.8% and 43.1% of people belongs to xx type 1 and type 2 respectively xxxx
* Column name convention: parent feature category; sub feature category; feature name
* Format: percentage form, float type
* Overview of XX dataset
* graph
* Overview of YY dataset
* graph

Features – person data

* Various data type: int, float, date, null, Boolean
* Unlike econ data, person data does not have category information in their column name
* Overview of xx data
* graph
* Overview of yy data
* graph

Data Quality

|  |  |
| --- | --- |
| Representability | Current features may not be enough for a comprehensive picture  (e.g. pensioners may want to take transfer but cannot get the approval due to regulatory reasons: the place they transfer to is on the prohibited list. Current features cannot show this information) |
| Completeness | Have missing values |
| Timeliness |  |
| Uniqueness | No duplicate record  (artificially duplicating records is NOT allowed) |
| Consistency | Abc dataset is set as default dominant dataset. When merging tables, data from non dominant dataset will always be replaced by the dominant one if they are referring to the same data |

Data Preprocessing

Dataset from various data sources will be merged into a single dataset in gold database

Features will be splitted into 4 sets: numerical, categorical, numerical to be binned, and others

* Econ features with “|” symbol are treated as numerical
* Econ features without “|” symbol are treated as categorical
* Exceptions specified by users
* Remaining features

Domain knowledge drop

* Drop date features
* Drop unique identifier features e.g. id
* Drop metadata
* Drop postcode as it is transformed into new feature
* Drop any obviously irrelevant data based on domain knowledge

Handling missing values

* Default null threshold is 40%
* Columns with more than 60% null values are dropped
* For columns with less than 10% null values, records with null will be dropped
* For columns with 10-40% null values, null values will be replaced by value with the highest frequency

Feature engineering

* New feature “in\_london” is created based on postcode feature (different behavior of people in London vs non London)

Data normalization

* Scaling standardization
* Min-max normalization

Encoding

* Target encoding, xx feature encoded as target variable
* Boolean to 1/0
* Binary features to 1/0
* Person data will be one hot encoded
* Data specifically specified by user will be binned and the one hot encoded instead

Dimensional reduction

Feature filtering for numerical features

* Variance filtering, default threshold 0.01
* Correlation filtering, default correlation threshold 0.85

Feature selection for numerical features

Random Forest Classifier to select top 10 features based on feature score

Data Security and Privacy

Data is depersonalized before being used. No personally identifiable information is included

Because of compliance/regulatory concerns, features related to Ethnicity and Religion are prohibited and excluded

## Data Dictionary

| **Column Name** | **Data Type** | **Description** | **Values/Range** | **Missing %** | **Notes** |
| --- | --- | --- | --- | --- | --- |
| transaction\_date | datetime | Date and time of transaction | 01/11/2022 00:22 to 02/11/2022 11:52 | 0% | UTC timezone assumed |
| customer\_id | int | Unique customer identifier | 1-250 | 0% | 250 unique customers |
| amount | float64 | Transaction amount in USD | 0.01-2491.44 | 0% | 2 decimal precision |
| date\_in\_sec | int64 | Unix timestamp in seconds | 1367-129165 | 0% | Derived field |
| No.\_of\_day | int64 | Day number in dataset | 1-2 | 0% | 1=Nov 1, 2=Nov 2 |
| merchant\_id | int64 | Unique merchant ID | 1-30 | 0% | 30 unique merchants |
| pos\_id | float64 | Point of sale terminal ID | 2-115 | 75% | Critical missing data |
| merchant\_established\_date | object | Merchant founding date | 26/06/1901 to 24/10/2020 | 0% | DD/MM/YYYY format |
| MCCs | int64 | Merchant category code | 3000, 5411, 5944 | 0% | 3 unique MCCs |
| merchant\_category | object | MCC description | "Airlines", "Grocery Stores/Supermarkets", "Jewelry Stores" | 0% |  |
| IP\_address | object | Originating IP address | Various | 40% | IPv4 format |
| type\_of\_credit\_card\_used | object | Card product type | "Premium Credit Card", "Store Credit Card", "Cash Back Card" | 0% | 3 card types |
| card\_present\_or\_not | object | Card presence | "card present", "card not present" | 0% | Key fraud indicator |
| is\_fraud | int64 | Fraud flag | 0 (legit), 1 (fraud) | 0% | Target variable |
| store\_card\_merchant\_id | float64 | Store card program ID | 1-28 | 85% | Mostly missing |

## Data Quality

**Credit Fraud Data Quality Assessment**  
*Generated: 20/Jun/2025*

1. Completeness Score: 72/100

**Missing Data Analysis:**

| **Column** | **Missing Count** | **% Missing** | **Impact** |
| --- | --- | --- | --- |
| pos\_id | 1,125 | 75.0% | High - Limits terminal analysis |
| IP\_address | 600 | 40.0% | Medium - Affects geo-analysis |
| store\_card\_merchant\_id | 1,275 | 85.0% | Low - Auxiliary field |

2. Consistency Check

**Issues Found:**

1. 3 transactions with amount = $0.00 (should be >0)
2. 7 transactions with future dates (11/3/2022)
3. 2 MCC codes not in standard list (5999, 3001)

3. Validity Assessment

**Field Validation Results:**

* **Amounts**: 99.8% valid (3 negative values found)
* **Dates**: 99.5% valid (7 future dates)
* **IPs**: 94% valid format (6% malformed)

4. Recommendations

**Critical Fixes (P0):**

1. Implement POS terminal data collection
2. Validate transaction amounts > $0
3. Fix date timezone issues

**Important Fixes (P1):**

1. Standardize MCC code list
2. Improve IP address capture

**Nice-to-Have (P2):**

1. Store card program documentation
2. Merchant metadata enrichment

## IDA/EDA Summary

*1. Key Findings*

***Fraud Prevalence***

* *1.62% fraud rate (24 fraud cases out of 1,500 transactions)*
* *Highly imbalanced dataset requiring special modeling techniques*

***Transaction Patterns***

* *Fraudulent transactions are significantly smaller (avg $27.15 vs $412.32)*
* *91.7% of fraud occurs in "card not present" transactions*
* *Fraud amounts often include unusual decimal values (e.g., $19.279230488928754)*

***Temporal Analysis***

* *Two distinct fraud peaks:*
  + ***Night****: 2-4 AM (38% of fraud cases)*
  + ***Morning****: 10-11 AM (29% of fraud cases)*
* *Average time between fraudulent transactions from same customer: 14 minutes*

***High-Risk Categories***

1. ***Jewelry Stores (MCC 5944)***
   * *4.2% fraud rate (3x overall average)*
   * *Average fraud amount: $312.45*
2. ***Airlines (MCC 3000)***
   * *45% of all fraud cases*
   * *Primarily "card not present" transactions*

***Geographical Indicators***

* *Top 3 suspicious IP ranges:*
  1. *222.222.x.x (78% fraud rate)*
  2. *111.111.x.x (65% fraud rate)*
  3. *196.89.x.x (53% fraud rate)*

*2. Recommendations*

***Immediate Actions***

1. *Implement real-time alerts for:*
   * *≥3 transactions from same customer in <30 minutes*
   * *Jewelry store transactions >$300 without card present*
   * *Transactions from 222.222.x.x IP range*
2. *Enhance verification for:*
   * *All airline transactions without card present*
   * *Transactions between 2-4 AM with amounts <$50*

***Strategic Initiatives***

1. *Develop machine learning model using:*
   * *Transaction amount*
   * *Time since last transaction*
   * *MCC code*
   * *IP reputation*
   * *Card presence*
2. *Data quality improvements:*
   * *POS terminal ID collection*
   * *IP address completeness*
   * *Store card program mapping*

***Risk Mitigation ROI***

| ***Control*** | ***Estimated Fraud Reduction*** | ***Implementation Cost*** |
| --- | --- | --- |
| *IP Blocking* | *32%* | *Low* |
| *MCC-Based Rules* | *28%* | *Medium* |
| *Velocity Checks* | *41%* | *High* |

## Preprocessing

## Data Security and Privacy